What is Data Science?

Michael L. Brodie, Computer Science and Artificial Intelligence Laboratory, MIT

Abstract

Data Science, a new discovery paradigm, is potentially one of the most significant advances of the early 21st century. Originating in scientific discovery, it is being applied to every human endeavor for which there is adequate data. While remarkable successes have been achieved, even greater claims have been made. Benefits, challenge, and risks abound. The science underlying *data science* has yet to emerge. Maturity is more than a decade away. This claim is based firstly on observing the centuries-long developments of its predecessor paradigms – empirical, theoretical, and Jim Gray's *Fourth Paradigm of Scientific Discovery* (Hey, Tansley & Tolle, 2009) (aka eScience, data-intensive, computational, procedural); and secondly on my studies of over 150 data science use cases, several data science-based startups, and, on my scientific advisory role for Insight[1], a Data Science Research Institute (DSRI) that requires that I understand the opportunities, state of the art, and research challenges for the emerging discipline of data science. This chapter addresses essential questions for a DSRI: *What is data science*? and *What is world-class data science research*? A companion chapter *On Developing Data Science* (Brodie, 2018b) addresses the development of data science applications and of the data science discipline itself.

1 Introduction

What can data science do? *What characteristics distinguish data science from previous scientific discovery paradigms*? *What are the methods for conducting data science*? *What is the impact of data science*? This chapter offers initial answers to these and related questions. A companion chapter (Brodie, 2018b) addresses the development of data science as a discipline, as a methodology, as well as data science research and education. Let's start with some slightly provocative claims concerning data science.

Data science has been used successfully to accelerate discovery of probabilistic outcomes in many domains. Piketty's (2014) monumental result on wealth and income inequality was achieved through data science. It used over 120 years of sporadic, incomplete, observational economic data, collected over ten years from all over the world (Brodie, 2014b). What is now called *computational economics* was used to establish the correlation, with a very high likelihood (0.90), that wealth gained from labor could never keep up with wealth gained from assets. What made front page news worldwide was a second, more dramatic correlation that there is a perpetual and growing wealth gap between the rich and the poor. This second correlation was not derived by data analysis but is a human interpretation of Piketty's data analytic result. It contributed to making *Capital in the 21st Century* the best-selling book on economics, but possibly the least read. Within a year, the core result was verified by independent analyses to a far greater likelihood (0.99). One might expect that further confirmation of Piketty's finding would be newsworthy; however, it was not as the more dramatic rich-poor correlation, while never analytically established had far greater appeal. This illustrates the benefits and risks of data science.

Frequently, due to the lack of evidence, economic theories fail. Matthew Weinzierl, a leading Harvard University economist, questions such economic modelling in general saying, "that the world is too complicated to be modelled with anything like perfect accuracy" and "Used in isolation, however, it can lead to trouble" (Economist, February 2018). Reputedly, Einstein said: "Not everything that counts can be counted. Not everything that's counted, counts". The hope is that data science and computational economics will provide theories that are fact-based rather than based on hypotheses of "expert" economists (Economist, January 2018) leading to demonstrably provable economic theories, i.e., what really happened or will happen. This chapter suggests that this hope will not be realized this year.

[1] https://www.insight-centre.org/

Many such outcomes[2] have led to verified results through methods outside data science. Most current data analyses are domain specific, many even specific to classes of models, classes of analytical methods, and specific pipelines. Few data science methods have been generalized outside their original domains of application, let alone to all domains (to illustrated in a moment). A rare and excellent exception is a generic scientific discovery method over scientific corpora (Nagarajan et. al., 2015) generalized from a specific method over medical corpora developed for drug discovery (Spangler et. al., 2014) that is detailed later in the chapter.

It is often claimed that data science will transform conventional disciplines. While transformations are underway in many areas, including supply chain management[3] (Waller and Fawcett, 2013) and chemical engineering (Data Science, 2018), only time and concrete results will tell the extent and value of the transformations. The companion chapter *On Developing Data Science* (Brodie, 2018b) discusses with the transformation myth.

While there is much science in many domain-specific data science activities, there is little fundamental science that is applicable across domains. To warrant the designation *data science*, this emerging paradigm requires fundamental principles and techniques applicable to all relevant domains, just as the scientific principles of the scientific method apply across many domains. Since most data science work is domain specific, often model- and method-specific, data science does not yet warrant the designation as a science.

This chapter explores the current nature of data science, its qualitative differences with its predecessor scientific discovery paradigms, its core value and components that, when mature, would warrant the designation *data science*. Descriptions of large-scale data science activities referenced in this chapter apply, scaled down, to data science activities of all sizes, including increasingly ubiquitous desktop data analytics in business.

2 What is data science?

Due to its remarkable popularity, there is a plethora of descriptions of data science, for example:

> *Data Science is concerned with analyzing data and extracting useful knowledge from it. Building predictive models is usually the most important activity for a Data Scientist*[4].

> *Data Science is concerned with analyzing Big Data to extract correlations with estimates of likelihood and error*. (Brodie, 2015a)

> *Data science is an emerging discipline that draws upon knowledge in statistical methodology and computer science to create impactful predictions and insights for a wide range of traditional scholarly fields*[5].

Due to data science being in its infancy, these descriptions reflect some of the many contexts in which it is used. This is both natural and appropriate for an emerging discipline that involves many distinct disciplines and applications. A definition of data science requires the necessary and sufficient conditions that distinguish it from all other activities. While such a definition is premature, a working definition can be useful for discussion. The following definition is intended to explore the nature of this remarkable new

---

[2] Not Piketty's, since computational economics can find *what* might have happened - patterns, each with a given likelihood - but lacks the means of establishing causal relationships, i.e., establishing *why*, based solely on observational data.

[3] Selecting the best delivery route for 25 packages from 15 septillion alternatives, an ideal data science application, may explain the some of the $1.3trn to $2trn a year in economic value projected to be gained in the transformation of the supply chain industry due to AI-based data analytics (Economist, March 2018).

[4] Gregory Piatetsky, KDnuggets, https://www.kdnuggets.com/tag/data-science

[5] Harvard Data Science Initiative https://datascience.harvard.edu

discovery paradigm. It is based on studying over 150 data science use cases and benefits from three years research and experience over a previous version (Brodie, 2015a). Like many data science definitions, it will be improved over the next decade in which data science will mature and gain the designation as a new science.

> *Data Science is* a *body of principles and techniques for applying data analytic methods to data at scale, including volume, velocity, and variety, to accelerate the investigation of phenomena represented by the data, by acquiring data, preparing and integrating it, possibly integrated with existing data, to discover correlations in the data, with measures of likelihood and within error bounds. Results are interpreted with respect to some predefined (theoretical, deductive, top-down) or emergent (fact-based, inductive, bottom-up) specification of the properties of the phenomena being investigated.*

A simple example of a data science analysis is the pothole detector developed at MIT (Eriksson et. al., 2008) to identify potholes on the streets of Cambridge, MA. The data was from inexpensive GPS and accelerometer devices placed in a fleet of taxis that drive over Cambridge streets. The model was designed *ad hoc* for this application. A model consists of the features (i.e., variables) essential to the analysis and the relationships amongst the features. It was developed in this case *ad hoc* by the team iteratively refining the model through imagination, observation, and analysis. Ultimately, it consisted of a large number of movement signatures, i.e., model features, each designed to detect specific movement types that may indicate potholes and non-potholes, e.g., manholes, railroad tracks[6], doors opening and closing, stopping, starting, accelerating, etc. Additionally, the size of the pothole was estimated by the size of the movement. The analytical method was the algorithmic detection and filtering of non-pothole signatures leaving as a result those movements that correlate with potholes with an estimated severity, likelihood, and error bound. The severity and likelihood estimates were developed *ad hoc* based on verifying some portion of the detected movements with the corresponding road surfaces thus contributing to estimating the likelihood that the non-potholes were excluded, and potholes were included. Error bounds were based on the precision of the equipment, e.g., motion device readings, network communications, data errors, etc. The initial result was many thousands of locations with estimated severities, likelihoods, and error bounds. Conversion of likely pothole locations (correlations) to actual potholes severe enough to warrant repair (causal relationships between movements and potholes) were estimated by a manual inspection of some percentage of candidate potholes. The data from the inspection of the actual locations, called ground truth, was used to verify the likelihood estimates and establish a threshold above which confidence in the existence of a pothole warranted sending out a repair crew to repair the pothole. The customer, the City of Cambridge, MA, was given a list of these likely potholes.

The immediate value of the pothole detector was that it reduced the search for potholes from manually inspecting 125 miles of roads and relying on citizen reports that takes months, to discovering

---

[6] The pothole models consist of a number of signature movements, i.e., abstractions used to represent movements of the taxi, only some of which are related to the road surface. Each signature movement was created using the data (variables or features) available from a smartphone including the clock for time, the GPS for geographic location (latitude and longitude), and the accelerometer to measure changes in velocity along the x, y, and z axes. For example, the taxi crossing a railroad track would result in many signature "single tire crossing single rail line" movements, one for each of four tires crossing each of several rail lines. A "single tire crossing single rail line" involves a sudden, short vertical (x-axis) acceleration combined with a short lateral (y-axis) movement, forward or backward, with little or no lateral (z-axis) movement. Discounting the railroad crossing as a pothole involves recognizing a large number of movements as a taxi is crossing a rail line - all combinations of "single tire crossing single rail line" forward or backward, at any speed, and at any angle - to determine the corresponding staccato of the multiple single tire events over multiple lines. The pothole model is clearly *ad hoc*, in contrast to well established models in physics and retail marketing.

likely, sever potholes within days of their creation. Since 2008, pothole detectors have been installed on city vehicles in many US cities. The pothole detector team created Cambridge Mobile Telematics that develops applications for vehicles sensor data, e.g., they annually produce reports on distracted driving across the USA based on data from over 100 million trips (Cambridge Mobile Telematics, 2018). While these applications were used initially by insurance companies they are part of the burgeoning domain of autonomous vehicles and are being used by the US National Academy of Sciences (Dingus T.A., 2016) for driving safety.

3 Data science is a new paradigm of discovery

Data science emerged from, and has many commonalities with, its predecessor paradigm, the scientific method[7]; however, they differ enough for data science to be considered a distinct, new paradigm. Like the scientific method, data science is based on principles and techniques required to conduct discovery activities that are typically defined in terms of a sequence of steps, called a workflow or pipeline; results are specified probabilistically and with error bounds based on the data, the model, and the analytical method used; and the results are interpreted in terms of the hypothesis being evaluated, the model, the methods, and the probabilistic outcome relative to the accepted requirements of the domain of the study. In both paradigms, *models* are collections of features (represented by variables that determine the data to be collected) that characterize the essential properties of the phenomenon being analyzed. Data corresponding to the features (variables) in the model are collected from real instances of the phenomena and analyzed using analytical methods developed for the type of analysis to be conducted and the nature of the data collected, e.g., different methods are required for integers uniformly distributed in time versus real numbers skewed due to properties of the phenomenon. The outcomes of the analysis are interpreted in terms of the phenomena being analyzed within bounds of precision and errors that result from the data, model, and method compared with the precision required in the domain being analyzed, e.g., particle physics requires precision of six standard deviations (six sigma). Data science differs paradigmatically from the scientific method in data, models, methods, and outcomes, as described below. Some differences may be due to data science being in its infancy, i.e., models for real-time cyber-attacks may not yet have been developed and proven; however, some differences, discussed below, are inherent. We are in the process of learning which is which.

3.1 Data science data, models, and methods

Data science data is often obtained with limited knowledge of the conditions under which the data was generated, collected, and prepared for analysis, e.g., data found on the web; hence, it cannot be evaluated as in a scientific experiment that requires precise controls on the data. Such data is called observational. Compared with empirical scientific data, data science data is typically, but not necessarily, at scale by orders of magnitude in one or more of volume, velocity, and variety. Scale requires management and analytic methods seldom required in empirical science.

Data science models used in most scientific domains have long histories of development, testing, and acceptance, e.g., the standard model of particle physics[8] emerged in 1961 after decades of development and has matured over the subsequent decades. In contrast, currently data science models, e.g., for real-time bidding for online advertising, are created on demand for each data science activity using many different, innovative, and *ad hoc* methods. Once a model is proven, they can be accepted and

[7] The scientific method is a body of techniques for investigating phenomena, acquiring new knowledge, or correcting and integrating previous knowledge. To be termed scientific, a method of inquiry is commonly based on empirical or measurable evidence subject to specific principles of reasoning. https://en.wikipedia.org/wiki/Scientific_method

[8] The Standard Model of particle physics is the theory describing three of the four known fundamental forces (the electromagnetic, weak, and strong interactions, and not including the gravitational force) in the universe, as well as classifying all known elementary particles. It was developed in stages throughout the latter half of the 20th century, through the work of many scientists around the world. https://en.wikipedia.org/wiki/Standard_Model

put into productive use with periodic tuning, e.g., real-time ad placement products. It is likely that many proven data science models will emerge as data science modelling matures. StackAdapt.com has developed such a model for Real-time Bidding and programmatic ad purchasing (RTB) that is its core capability and intellectual property with which it has become a RTB world leader amongst 20 competitors worldwide. The StackAdapt model is used to scan 10 BN data points a day and manage up to 150,000 ad opportunity requests per second during peak times.

Data science analytical methods, like data science models, are often domain- and data-specific and are developed exclusively for a specific data science activity. There are generic methods, often named by a class name. For example, the primary classes of Machine Learning algorithms[9] are: Linear Classifiers: Logistic Regression, Naive Bayes Classifier; Support Vector Machines; Decision Trees; Boosted Trees; Random Forest; Neural Networks; and Nearest Neighbor. There are generic algorithms for each class each of which can be applied in many domains. However, to be applied in a specific use case they must be refined or tuned often to the point of being applicable in that use case only. This is addressed in the next section that questions whether there are, as yet, underlying, thus generalizable, principles in data science..

Both models and methods require tuning or adjusting in time as more knowledge and data are obtained. Empirical scientific models tend to evolve slowly, e.g., the standard model of particle physics is modified slowly[10]; in contrast, data science models typically evolve rapidly throughout their design and development, and even in deployment, using dynamic learning. Typically, models and methods are trained using semi-automatic methods by which specific data or outcomes, called ground truth, are confirmed by humans as real to the model or method. More automatic methods, e.g., reinforcement learning and meta-learning[11], are being developed by which models and methods are created automatically (Silver, 2017).

3.2 Data science fundamentals: Is data science a science?

Currently, most data science results are domain-, method-, and even data-specific. This raises the question as to whether data science is yet a science, i.e., with generalizable results, or merely a collection of sophisticated analytical methods, with, as yet, a few underlying principles emerging, such as Bayes' Theorem, Uncle Bernie's rule[12], and Information Bottleneck theory. The scientific method is defined by principles that ensure scientific objectivity, such as empirical design and the related controls to govern experimental design and execution. These and other scientific principles make experiments "scientific", the minimum requirement for a result to be considered scientific. Scientific experiments vary across domains, such as the statistical significance required in a given domain, e.g., two sigma has traditionally been adequate in many domains besides particle physics. A necessary, defining characteristic of data science is that the data is either at scale (Big Data) or observational (collected without knowing the provenance - what controls were applied or with no controls uniformly applied) as is generally the case in economics and social sciences. Under those conditions, data science cannot be "scientific", hence accommodations must be made to draw conclusions from analysis over such data. As data science is just emerging in each domain, we have few principles or guidelines per domain, e.g., statistical significance of results, or across all domains, e.g., the extent to which statistical significance is required in any data science analysis. The above mentioned pothole analysis was designed and executed by sheer intuition beyond the general ideas of identifying the hypothesis (find potholes using motion devices in taxis), experimental design (put devices in taxis and record their signals), modeling (what features are critical), and analysis (what motions indicate potholes, and which do not), and iteration of the model and analysis until acceptable precision was reached. The pothole data science activity did not draw on previous methods, nor did it offer, i.e., was not cited, principles for modeling, methods, or process.

---

[9] https://medium.com/@sifium/machine-learning-types-of-classification-9497bd4f2e14

[10] Validating the Higgs-Boson took 49 years.

[11] http://bair.berkeley.edu/blog/2017/07/18/learning-to-learn/

[12] See Morgan, N., & Bourlard, H. (1990). Generalization and parameter estimation in feedforward nets: Some experiments. In Advances in neural information processing systems (pp. 630-637).

Another practical example is at Tamr.com that offers one of the leading solutions for curating or preparing data at scale, e.g., data from 100,000 typically heterogeneous data sources. It launched initially with a comprehensive solution in the domain of information services. Tamr soon found that every new domain required a substantial revision of the machine learning component. Initially, like most AI-based startups, their initial solution was not generalizable. As can be seen at Tamr.com, Tamr now has solutions in many domains for which they have substantial commonality in the underlying solutions.

Another fundamental difference between science and data science concerns the *scale and nature of the outcomes.* The scientific method is used to discover causal relationships between a small number of variables that represent the essential characteristics of the natural phenomena being analyzed. The experimental hypothesis defines the correlation to be evaluated for causality. The number of variables in a scientific experiment is kept small due to the cost of evaluating a potentially vast number of combinations of variables of interest. PhD theses, i.e., an experiment conducted by one person, are awarded on experiments with two or three but certainly less than ten variables. Large-scale experiments, e.g. LIGO[13], Kepler[14], and Higgs-Boson, may consider 100s of variables and take years and thousands of scientists to evaluate. Determining whether a correlation between variables is causal tends to be an expensive and slow process.

Data science, on the other hand, is used to rapidly discover as many correlations between the data values as exist in the data set being analyzed, even with very large models (millions of variables) and vast data sets. Depending on the analytical method used, the number of variables in a data science analysis can be effectively unlimited, e.g., millions, even billions, as can be the number of correlations between those variables, e.g., billions or trillions. Data science analytics are executed by powerful, efficient algorithms using equally powerful computing infrastructure (CPUs, networks, storage). The combined power of new algorithms and infrastructure in the 1990's led to the current efficacy of machine learning that in turn contributed to the emergence of data science.

3.3 The prime benefit of data science is accelerating discovery

Data science and empirical science differ dramatically, hence paradigmatically, in the scale of the data analyzed. Scientific experiments tend to evaluate a small number, e.g., 10s or 100s, of correlations to determine if they are causal, and do so over long periods of time, e.g., months or years. In contrast, data science can identify effectively unlimited numbers of correlations, e.g., millions, billions, or more, in short time periods, from minutes to days. It is in this sense that data science is said to *accelerate discovery.* Originally developed in the 1990's for scientific discovery, the remarkable results of data science have resulted in its being applied to all endeavors for which adequate data is available. *The prime benefit of data science is that it is a new paradigm for accelerating discovery*, in general.

Ideally, data science is used to accelerate discovery by rapidly reducing a vast search space to a small number of correlations that are likely to be casual, as indicated by their estimated probability. Depending on the resources available, some number of the probabilistic correlations are selected to be analyzed for causality by well-established (non-data science) means in the domain being analyzed. For example, data science has been used to accelerate cancer drug discovery. The Baylor-Watson study (Spangler et. al., 2014) used data science methods to identify nine likely cancer drug candidates. It used a simple, novel method to further evaluate their likelihood. The original analysis was conducted over drug research results published up to 2003 and identified nine likely candidate drugs. The likelihood of those nine candidate drugs was raised significantly when the research published from 2003 to 2013 showed that seven of the nine candidates had been validated as genuine cancer drugs. This raised the likelihood that the remaining two candidate drugs were real. Standard EPA-approved drug development and clinical trial testing were then used to develop the two new drugs. In this case, data science accelerated drug discovery for a specific type of cancer. It started with a vast search space of cancer research results from 240,000 papers. In three months it discovered the two highly likely cancer drug candidates. Conventional

[13] http://www.ligo.org/

[14] https://keplerscience.arc.nasa.gov/

cancer drug discovery typically discovers one drug every two to three years. These times do not include the drug development and clinical trial periods.

3.4 Causal reasoning in data science is complex and can be dangerous

Just as the scale is radically different so is the nature of the results. The scientific method discovers results that, if executed correctly, are definitive, i.e., true or false, with a defined probability and error bound, that a hypothesized relationship is causal. Data science discovers a potentially large number of correlations each qualified by a probability and error bound that indicate the likelihood that the correlation may be true. *Data science is used to discover correlations; it is rarely used to determine causal relationships*. The previous sentence is often misunderstood not just by novices, but also, unfortunately, by data scientists. Empirical science discovers causal relationships in one step. Data science is frequently used to discover causal relationships in two steps: First, discover correlations with a strong likelihood of being causal; then use non-data science methods to validate causality.

Causality is the Holy Grail of science, scientific discovery, and if feasible, of data science. Typically, the goal of analyzing a phenomenon is to understand ***Why*** some aspects of the phenomenon occur, for example, why does it rain? Prior to a full understanding of the phenomenon, initial discovery is often used to discover ***What*** conditions prevail when the phenomenon manifests, e.g., as rain starts and during rain many raised umbrellas can be observed. A more informed observer may also discover specific climatic conditions. All of the conditions observed to be present consistently before and during the rain could be said to be correlated with rain. However, correlation does not imply causation, e.g., raised umbrellas may be correlated with rain, but do not cause the rain (Brodie, 2014a). A more realistic example comes from an online retailer that observing that increased sales were correlated with customers purchasing with their mobile app, invested significantly to get their app onto many customers' smartphones. However, the investment was lost since sales did not increase. Increased purchases were correlated with mobile apps on customers' smartphones; however, the causal factor was customer loyalty and, due to their loyalty, most loyal customers already had the app on their smartphones.

Data Science is used predominantly to discover ***What***. Empirical science and many other methods are used to discover ***Why*** (Brodie, 2018a). Data science is often used to rapidly reduce the search space from a vast number of correlations or possible results to a much smaller number. The much smaller number of highly probable results are then analyzed with non-data science methods, such as scientific experiments or clinical trials, to verify or reject the result, i.e., automatically generated hypotheses, as causal.

There are mathematics and methods claimed for deducing causal effects from observational data (i.e., data not from controlled experiments but from surveys, censuses, administrative records, and other typically uncontrolled sources such as in Big Data and data science). They are very sophisticated and require a deep understanding of the mathematics, statistics, and related modelling methods. Judea Pearl has developed such methods based on statistics, Bayesian networks, and related modelling, see (Pearl, 2009a,b,c). For decades, statisticians and econometricians have developed such methods with which to estimate causal effects from observational data, since most social and economic data is purely observational (Winship et. al., 1999).

Causal reasoning involves going beyond the mathematics and modelling for data science in which correlations are obtained. "One of Pearl's early lessons is that it's only possible to draw causal conclusions from observational (correlational) data if you are willing to make some assumptions about the way that the data were sampled and about the absence of certain confounding influences. Thus, my understanding is that one can draw causal conclusions, but it's important to remember that these are really conditional on the validity of those assumptions." says Peter Szolovits, Professor, CSAIL, MIT, with a decade of experience applying data science in medical contexts for which he provided an example[15].

[15] The full quote from personal communication: "There are various sophisticated ways to do all this but let me give you a relatively simple example: Suppose that we observe that in some cohort of patients, some were treated with drug X and others with drug Y. Suppose further that we see that fewer of the X patients died than of the Y ones. It's certainly NOT acceptable to conclude that X is a better drug, because we can't exclude the possibility that the treating doctors' choice of X or Y depended on some characteristics of

Finding correlations between variables in (Big) data together with probabilities or likelihoods of the correlation occurring in the past or future, are relatively easy to understand and safe to report. Making a causal statement can be misleading or dangerous depending on the proposed actions to be taken as a consequence. Hence, I do not condone nor confirm causal reasoning; it is above my pay grade; hence, I quote experts on the topic rather than make my own assertions. I recommend that causal reasoning not be applied without the required depth of knowledge and experience, because making causal statements as a result of data science analysis could be dangerous. In lecturing on correlation versus causation for over five years, I have found that an inordinate amount of interest is given to this difficult and little understood topic, perhaps with a desire to be able to provide definitive answers, even when there are none. I have found no simple explanation. You either study, understand, and practice causal reasoning with the appropriate care or simply stay away until you are prepared. Experts are appropriately cautious. "I have not, so far, made causal claims based on my work, mainly because I have not felt strongly enough that I could defend the independence assumptions needed to make such claims. However, I think the kinds of associational results are still possibly helpful for decision makers when combined with intuition and understanding. Nevertheless, I think most clinicians today do not use predictive models other than for more administrative tasks such as staffing or predicting bed occupancy" – Peter Szolovits, MIT. "I firmly believe that [deriving] causal results from observational data is one of the grand challenges of the data science agenda!" – David Parkes, co-lead of the Harvard Data Science Initiative. "Pearl once explained those ideas to me personally at Santa Catalina workshop, but I still don't fully understand them either :)" – Gregory Piatetsky-Shapiro, President of KDnuggets, co-founder of KDD Conferences and ACM SIGKDD.

3.5 Data science flexibility: data-driven or hypothesis-driven

Empirical science and data science have another fundamental difference. The scientific method uses deductive reasoning, also called hypothesis-driven, theory-driven, and top-down. Deductive reasoning is used when specific hypotheses are to be evaluated against observations or data. A scientific experiment starts by formulating a hypothesis to be evaluated. An experiment is designed and executed, and the results interpreted to determine if the hypothesis is true or false under the conditions defined for the hypothesis. It is called theory-driven in that a theory is developed, expressed as a hypothesis, and an experiment designed to prove or invalidate the hypothesis. It is called top-down since the experiment starts at the top – with the idea – and goes down to the data to determine if the idea is true.

Data science can be hypothesis-driven. That is, as with empirical science, a data science activity can start with a hypothesis to be evaluated. Unlike empirical science, the hypothesis can be stated with less precision and the models, methods, and data can be much larger in scale, i.e., more variables, data volume, velocity, and variety. In comparison, data science accelerates discovery by rapidly reducing a vastly larger search space than would have been considered for empirical methods, to a small set of likely correlations; however, unlike empirical science, the results are correlations that require additional, non-data science methods to achieve definitive, causal results.

One of the greatest advantages of data science is that it can discover patterns or correlations in data at scale vastly beyond human intellectual, let alone temporal, capacity; far beyond what humans

---

the patient that also influenced their likelihood of survival. E.g., maybe the people who got Y were much sicker to start with, because Y is a stronger and more dangerous drug, so it is only given to the sickest patients.

One way to try to mitigate this is to build a model from all the data we have about the patients in the cohort that predicts whether they are likely to get X or Y. Then we stratify the cohort by the probability of getting X, say. This is called a *propensity score*. Among those people with a high score, most will probably actually get X (that's how we built the model), but some will nevertheless get Y, and *vice versa*. If we assume that the doctors choosing the drug have no more information than the propensity model, then we treat their choice to give X or Y as a random choice, and we analyze the resulting data as if, for each stratum, patients were randomized into getting either X or Y, as they might have been in a real clinical trial. Then we analyze the results under that assumption. For many of the strata where the propensity is not near .5, the drugs given will be unbalanced, which makes the statistical power of the analysis lower, but there are statistical methods for dealing with this. Of course, the conclusions one draws are still very much dependent of the assumption that, within each stratum, the doctors' choice of drug really is random, and not a function of some difference among the patients that was not captured in the data from which the propensity model was built.

This is just one of numerous methods people have invented, but it is typical of the kinds of assumptions one has to make in order to draw causal conclusions from data."

could have conceived. Of course, a vast subset of those found may be entirely spurious. Data science can use inductive reasoning, also called bottom-up, data-driven, or fact-based analysis, not to evaluate specific hypotheses but using an analytical model and method to identify patterns or correlations that occur in the data with a specific frequency. If the frequency meets some predefined specification, e.g., statistical significance in the domain being analyzed, it can be interpreted as a measure of likelihood of the pattern being real. As opposed to evaluating pre-defined hypotheses in the theory-driven approach, the data-driven approach is often said to "automatically" generate hypotheses, as in (Nagarajan, 2015). The inductive capacity of data science is often touted as its magic as the machine or methods such as machine learning, "automatically" and efficiently discover likely hypotheses from the data. While the acceleration and the scale of data being analyzed are major breakthroughs in discovery, the magic should be moderated by the fact that the discovered hypotheses are derived from the models and methods used to discover them. The appearance of magic may derive from the fact that we may not understand how some analytical methods, e.g., some machine learning and deep learning methods, derive their results. This is a fundamental data science research challenge as we would like to understand the reasoning that led to a discovery, as is required in medicine, and in 2018 in the European Union, by law (the General Data Protection Regulation (GDPR[16])).

3.6 Data science is in its infancy

The excitement around data science and its many successes are wonderful, and the potential of data science is great, but these positive signs can be misleading. Not only is data science in its infancy as a science and a discipline, its current practice has a large learning curve related largely to the issues raised above. Gartner, Forrester, and other technology analysts report that most (80%) early (2010-2012) data science projects in most US enterprises failed. In late 2016, Gartner reported that while most enterprises declare data science as a core expertise, only 15% claim to have deployed big data projects in their organization (Gartner, 2016). Analysts predict 80+% failure rate through 2017 (Demirkan & Dal, 2014) (Veeramachaneni, K. 2016) (Lohr & Singer, 2016).

3.7 It's more complicated than that

Data science methods are more sophisticated than the above descriptions suggest, and data-driven analyses are not as pure. Data science analytical methods and models do not discover any and all correlations that exist in the data since they are discovered using algorithms and models that incorporate some hypotheses that could be considered biases. That is, you discover what the models and methods are designed to discover. One must be objective in data science across the entire workflow - data selection, preparation, modelling, analysis, and interpretation; hence, a data scientist must always Doubt and Verify (Brodie, 2015b).

It may be useful to experiment with the models and methods. When a data science analysis reduces a vast search space, it (or the observing human) may learn something about the discovered correlations and may warrant an adjustment and re-running the model, the method, or even adjusting the data set. Hence, iterative learning cycles may increase the efficacy of the analysis or simply provide a means of exploring the data science analysis search space.

Top-down and bottom-up analytical methods can be used in combination, as follows. Start with a bottom-up analysis that produces N candidate correlations. Select a subset of K of the correlations with an acceptable likelihood and treat them as hypotheses to be evaluated. Then use them to run hypothesis-driven data science analyses and determine, based on the results, which hypotheses are again the most likely or perhaps even more likely than the previous run and discard the rest. These results can be used in turn to redesign the data science analysis, e.g., iteratively modify the data, model, and method, and repeat the cycle. This approach is used to explore data, models, and methods - the main components of a data science activity. This method of combining top-down and bottom-up analysis has been proposed by

[16] https://en.wikipedia.org/wiki/General_Data_Protection_Regulation

CancerCommons, as a method for accelerating the development of cancer cures as part of the emerging field of translational medicine.[17]

4 Data science components

Extending the analogy with science and the scientific method, data science, when mature, will be a systematic discipline with components that are applicable to most domains – to most human endeavors. There are four categories of data science components, all emergent in the data science context awaiting research and development: 1) principles, data, models, and methods; 2) data science pipelines; 3) data science infrastructure; and 4) data infrastructure. Below, we discuss these components in terms of their support of a specific data science activity.

Successful data science activities have developed and deployed these components specific to their domain and analysis. To be considered a science, these components must be generalized across multiple domains, just as the scientific method applies to most scientific domains, and in the last century has been applied to domains previously not considered scientific, e.g., economics, humanities, literature, psychology, sociology, and history.

4.1 Data science principles, data, models, and methods

A data science activity must be based on ***data science principles***, ***models***, and *analytical* ***methods***. Principles include those of science and of the scientific method applied to data science, for example, deductive and inductive reasoning, objectivity or lack of bias relative to a given factor, reproducibility, and provenance. Particularly important are collaborative and cross-disciplinary methods. How do scientific principles apply to discovery over data? What principles underlie evidence-based reasoning for planning, predicting, decision-making, and policy-making in a specific domain?

In May 2017, the *Economist* declared, on its front cover, that **data** was *The World's Most Valuable Resource* (Economist, May 2017). Without data there would be no data science or any of its benefits. Data management has been a cornerstone of computer science technology, education, and research for over 50 years, yet Big Data that is fueling data science, is typically defined as data at volumes, velocities, and variety that cannot be handled by data management technology. A simple example is that data management functions in preparing data for data analysis take 80% of the resources and time for most data science activities. Data management research is in the process of flipping that ratio so that 80% of resources can be devoted to analysis. Discovering data required for a data science activity whether inside or outside an organization is far worse. Fundamental data research is required in each step of the data science pipeline to realize the benefits of data science.

A data science activity uses one or more models. A model represents the parameters that are the critical properties of the phenomenon to be analyzed. It often takes multiple models to capture all relevant features. For example, the LIGO experiment, that won the 2017 Nobel Prize in Physics for empirically establishing the existence of Einstein's gravitational waves, had to distinguish movement from gravitational waves from seismic activity and 100,000 other types of movement. LIGO required a model for each movement type so as to recognize it in the data and discard it as gravitational wave activity. Models are typically domain specific, e.g., seismic versus sonic, and are often already established in the domain. Increasingly, models are developed specifically for a data science activity, e.g., feature extraction from a data set is common for many AI methods. Data science activities often require the continuous refinement of a model to meet the analytical requirements of the activity. This leads to the need for model management to capture the settings and results of the planned and evaluated model variations. It is increasingly common, as in biology, to use multiple, distinct models, called an ensemble of models, each of which provides insights from a particular perspective. Each model, like each person in Plato's Allegory of the Cave, represents a different perspective of the same phenomenon, what Plato called shadows. Each model – each person – observes what appears to be the same phenomenon, yet each sees it differently. No one model – person – sees the entire thing, yet collectively they capture the whole phenomenon from many perspectives. It may also be that a critical perspective is missed. It is rarely

[17] The National Center for Advancing Translational Sciences https://ncats.nih.gov

necessary, feasible, or of value to integrate different perspectives into a single integrated model. After all, there is no ultimate or truthful model save the phenomenon itself. Ensemble or shadow modelling is a natural and nuanced form of data integration (Liu, 2012) analogous to ensemble modelling in biology and ensemble learning (Dietterich, 2000) and forecasting in other domains.

A data science activity can involve many analytical methods. A given method or algorithm is designed to analyze specific features of a data set. There are often variations of a method depending on the characteristics of the data set, e.g., sparse or dense, uniform or skewed, data type, data volume, etc., hence methods must be selected, or created, and tuned for the data set and analytical requirements, and validated. In an analysis, there could be as many methods as there are specific features with corresponding specific data set types. Compared with analytical methods in science, their definition, selection, tuning, and validation in data science often involves scale in choice and computational requirements. Unless they are experts in the related methods, it is unlikely that a practicing data scientist understands the analytical method, e.g., a specific machine learning approach, that they are applying relative to the analysis and data characteristics, let alone the thousands of available alternatives. Anecdotally, I have found that many practicing data scientists use the algorithms that they were taught rather than selecting the one most applicable to the analysis at hand. There are significant challenges in applying sophisticated analytical models and methods in business (Forrester, 2015). Having selected or created and refined the appropriate model, i.e., collection of features that determine the data to be collected, collected and prepared the data to comply with the requirements of the model, and selected and refined the appropriate analytical method, the next challenge is interpreting the results and, based on the data, model, and method, evaluate the likelihood, within relevant error bounds, that the results are meaningful hypotheses worthy of validating by other means.

4.2 Data science workflows or pipelines

The central organizing principle of a data science activity is its ***workflow*** or ***pipeline*** and its life cycle management (NSF, 2016). A data science pipeline is an end-to-end sequence of steps from data discovery to the publication of the qualified, probabilistic interpretation of the result in the form of a data product. A generic data science pipeline, such as listed below, is comprehensive of all data science activities, hence can be used to define the *scope of data science*.

1. Raw data discovery, acquisition, preparation, and storage as curated data in data repositories
2. Selection and acquisition of curated data from data repositories for data analysis
3. Data analysis
4. Results interpretation
5. Result publication and optionally operationalize the pipeline for continuous analyses

The state of the art of data science is such that every data science activity has its own unique pipeline, as each data science activity is unique. Due to the emergence and broad applicability of data science, there is far more variation across data science pipelines than across conventional science pipelines. Data science will benefit, as it develops, from a better understanding of pipelines and guidance on their design and development.

Data science pipelines are often considered only in terms of the analytics, e.g., the machine learning algorithms used to derive the results in step 3. However, most of the resources required to design, tune, and execute a data science activity are required not for data analysis, steps 3 and 4 of a data science pipeline, but for the design and development of the pipeline and for steps 1 and 2.

The design, development, and tuning of an end-to-end pipeline for a data science activity typically poses significant data modelling, preparation, and management challenges often requiring significant resources and time required to develop and execute a data science activity. Two examples are astrophysical experiments, the Kepler Space Telescope launched in 2009 to find exoplanets and Laser Interferometer Gravitational-Wave Observatory (LIGO) that was awarded the 2017 Nobel Prize in Physics. Initial versions of the experiments failed not because of analysis and astrophysical aspects and models, but due to the data pipelines. Due to unanticipated issues with the data, the Kepler Science Pipeline had

to be rewritten (Jenkins, 2010) while Kepler was inflight retaining all data for subsequent corrected processing. Similarly, earth-based LIGO's pipeline was rewritten (Singh, 2007) and renamed Advanced LIGO. Tuning or replacing the faulty pipelines delayed both experiments by approximately one year.

Once the data has been acquired, the most time-consuming activity in developing a pipeline was data preparation. Early data science activities in 2003 reported 80-90% of resources devoted to data preparation (Dasu & Johnson, 2003). By 2014 this was reduced to 50-80% (Lohr, 2014). In specific cases, this cost negatively impacted some domains (Reimsbach-Kounatze, 2015) due to the massive growth of acquired data. As data science blossomed so did data volumes, leading experts in 2015 to analyze the state of the art and estimating that data preparation typically consumed 80% of resources (Castanedo, 2015). By then products to curate data at scale, such as Tamr.com, were maturing and being more widely adopted. Due to the visibility of data science, the popular press surveyed data scientists to confirm the 80% estimates (Press, 2016; Thakur 2016). In 2017, technical evaluations of data preparation products and their use again identified the 2003 estimates of 80% (Mayo, 2017) (Gartner G00315888, 2017).

4.3 Data science and data infrastructures

The core technical component for a data science activity is a ***data science infrastructure*** that supports the steps of the data science pipeline throughout its life cycle. A data science infrastructure consists of a workflow platform that supports the definition, refinement, execution, and reporting of data science activities in the pipeline. The workflow platform is supported by the infrastructure required to support workflow tasks such as data discovery, data mining, data preparation, data management, networking, libraries of analytical models and analytical methods, visualization, etc. To support user productivity, a user interface is required for each class of user, each with their own user experience. There are more than 60 such data science platforms - a new class of product - of which 16 meet analysts' requirements (Gartner G00301536, 2017) (Gartner G00326671, 2017) (Forrester, 2017). These products are complex with over 15 component products such as database management, model management, machine learning, advanced analytics, data exploration, visualization, and data preparation. The large number of products reflects the desire to get into a potentially large, emerging market; regardless of their current ability to support data science[18].

Data, the world's most valuable resource (Economist, May 2017), is also the most valuable resource for the data science activities of an organization (e.g., commercial, educational, research, governmental) and for entire communities. While new data is always required for an existing or new data science activity, data science activities of an organization require a ***data infrastructure*** – a sustainable, robust data infrastructure consisting of repositories of raw and curated data required to support the data requirements of the organization's data science activities with the associated support processes such as data stewardship. Many organizations are just developing data infrastructures for data science, aka data science platforms. The best known are those that support large research communities. The US National Research Foundation is developing the *Sustainable Digital Data Preservation and Access Network Partners* to support data science for national science and engineering research and education. The 1000 Genomes Project Consortium created the world's largest catalog of genomic differences among humans, providing researchers worldwide with powerful clues to help them establish why some people are susceptible to various diseases. There are more than ten additional genomics data infrastructures, including the Cancer Genome Atlas of the US National Institutes of Health, Intel's Collaborative Cancer Cloud, and the Seven Bridges Cancer Cloud. Amazon hosts[19] the 1000 Genome Project and 30 other public data infrastructures on topics such as geospatial and environmental datasets, genomics and life science datasets, and datasets for machine learning. The Swiss Data Science Center started developing the Renga platform[20] to support data scientists with their complete workflow.

---

[18] By 1983 in response to the then emerging technology of relational database management systems (DBMSs) there were over 100 Relational DBMSs of which five survived.

[19] https://aws.amazon.com/public-datasets/

[20] https://datascience.ch/renga-platform/

5 What is the method for conducting data science?

A data science activity is developed based on data science principles, models and analytical methods. The result of its design and development is a data science pipeline that will operate on a data science infrastructure, or platform, and will access data in a data infrastructure. There are a myriad of design and development methods to get from the principles to the pipeline. What follows is a description of a fairly generic data science method.

The ***data science method***, until better alternatives arise, is modelled on the scientific method. The following is one example of applying the empirical approach to data science analysis, analogous to experimental design for science experiments. Each step requires verification, e.g., using experts, published literature, previous analysis; and continuous iterative improvement to reach results that meet a predefined specification. Each step may require revisiting a previous step, depending on its outcome. As with any scientific analysis, every attempt should be made to avoid bias, namely, attempting to prove preconceived ideas beyond the model, methods, and hypotheses. The method may run for hours to days for a small analysis; months, as for the Baylor-Watson drug discovery (Spangler et. al., 2014); or years, as for the Kepler Space Telescope and LIGO. Design and development times can be similar to run times. Otto for example, a German e-commerce merchant, developed over months an AI-based system that predicts with 90% accuracy what products will be sold in the next 30 days and a companion system that automatically purchases over 200,000 products[21] a month from third-party brands without human intervention. Otto selected, modified, and tuned a deep-learning algorithm originally designed for particle-physics experiments at CERN (Economist, April 2017). These systems run continuously.

5.1 A Generic Data Science Method[22]

1. Identify the phenomena or problem to be investigated. What is the desired outcome?
2. Using domain knowledge, define the problem in terms of features that represent the critical factors or parameters to be analyzed (the WHAT of your analysis, that collectively form the model), based on the data likely to be available for the analysis. Understanding the domain precedes defining hypotheses to avoid bias.
3. If the analysis is to be top-down, formulate the hypotheses to be evaluated over the parameters and models.
4. Design the analysis in terms of an end-to-end workflow or pipeline from the data discovery and acquisition, through analysis and results interpretation. The analysis should be designed to identify probabilistically significant correlations (*What*) and set requirements for acceptable likelihoods and error bounds.
5. Ensure the conceptual validity of the data analysis design.
6. Design, test, and evaluate each step in the pipeline, selecting the relevant methods, i.e., class of relevant algorithms, in preparation for developing the following steps.
    a. Discover, acquire, and prepare data required for the parameters and models ensuring that the results are consistent with previous steps.
    b. For each analytical method, select and tune the relevant algorithm to meet the analytical requirements. This and the previous step are highly interrelated and often executed iteratively until the requirements are met with test or training data.
    c. Ensure the validity of the data analysis implementation.
7. Execute the pipeline ensuring that requirements, e.g., probabilities and error bounds, are met.
8. Ensure empirical (common sense) validation - the validity of the results with respect to the phenomena being investigated.

[21] Stock Keeping Units (SKUs).

[22] This set of steps was derived from analyzing over 150 data science activities. It's purpose is as a basis for guidance for those new to data science and as one alternative to data scientists looking for commonality across domains.

9. Interpret the results with respect to the models, methods, and data analytic requirements. Evaluate the results (patterns or correlations) that meet the requirements for causality to be validated by methods outside data science.
10. If the pipeline is to operate continuously, operationalize and monitor the pipeline and its results.

6 What is data science in practice?

Each data science activity develops its own unique data science method. Three very successful data science activities are described below in point form descriptions, using the above terminology to illustrate the components of data science in practice. They were conducted over 18, 20, and 2 years respectively. Their data science pipelines operated for 4 years, 3 years (to date), and 3 months respectively.

6.1 Kepler Space Telescope: Discovering Exoplanets

The Kepler Space Telescope, initiated in 1999, and its successor project K2, have catalogued thousands of exoplanets by means of data analytics over Big Data. A detailed description of Kepler and access to its data is at NASA's Kepler & K2 Website[23].

- **Objective and phenomenon**: Discover exoplanets in telescopic images
- **Project:** NASA-led collaboration of US government agencies, universities, and companies.
- **Critical parameters**: Over 100, e.g., planet luminosity, temperature, planet location relative to its sun.
- **Models**: There are over 30 established astrophysical models. A key Kepler model is the relationship between luminosity, size, and temperature. This model was established a century ago by Ejnar Hertzsprung and Henry Russell. This illustrates the fact that data science involves many models and analytical methods that have nothing to do with AI.
- **Methods**: Over 100, e.g., multi-scale Bayesian Maximum A Priori method used for systematic error removal from raw data. AI was not a principle method in this project.
- **Hypotheses** (stated in Kepler documents as a query): Five, including "Determine the percentage of terrestrial and larger planets that are in or near the habitable zone of a wide variety of stars".
- **Data**: 100's of data types described in the Data Characteristics Handbook[24] in the NASA Exoplanet Archive[25]
- **Pipeline**: The *Kepler Science Pipeline*[26] failed almost immediately after launch due to temperature and other unanticipated issues. After being repaired from earth, it worked well for 4 years.
- **Data discovery and acquisition**: Required approximately 90% of the total effort and resources.
- **Algorithm selecting and tuning**: Models and methods were selected, developed, tuned and tested for the decade from project inception in 1999 to satellite launch in 2009, and were refined continuously.
- **Verification:** Every model and method were verified, e.g., exoplanet observations were verified using the Keck observatory in Hawaii.
- **Probabilistic outcomes**[27]

  Kepler:
  - Candidates (<95%): 4,496
  - Confirmed (>99%): 2,330
  - Confirmed: <2X Earth-size in habitable zone: 30
  - Probably (<99%): 1,285
  - Probably not (~99%): 707

  K2:

[23] https://keplerscience.arc.nasa.gov/
[24] https://archive.stsci.edu/kepler/manuals/Data_Characteristics.pdf
[25] https://exoplanetarchive.ipac.caltech.edu/docs/KeplerMission.html
[26] https://keplerscience.arc.nasa.gov/pipeline.html
[27] Kepler's data is available at http://exoplanetarchive.ipac.caltech.edu

- Candidate (<95%): 521
- Confirmed (>99%): 140

6.2 LIGO: Detecting Gravitational Waves

The LIGO project detected cosmic gravitational waves predicted by Einstein's 1916 Theory of General Relativity for which it's originators were awarded the 2017 Nobel Prize. Project information and its data are available at the LIGO Scientific Collaboration website[28].

- **Objective and phenomenon**: Observe cosmic gravitational waves.
- **Project**: Initiated in 1997 with 1,000 scientists in 100 institutes across 18 countries.
- **Equipment**: Laser Interferometer Gravitational-Wave Observatory (world's most sensitive detector).
- **Go Live:** September 2015 (after a massive upgrade).
- **Data:** 100,000 channels of measurement of which one is for gravitational waves.
- **Models**: At least one model per channel.
- **Methods**: At least one data analysis method per data type being analyzed. Initially, AI was not used. In the past two years Machine Learning has been found to be very effective in many areas, e.g., detector malfunctions, earthquake detection.
- **Challenges:** Equipment and pipeline (as is typical in data science activities).
- **Results:**
    - In September 2015 (moments after reboot following the massive upgrade), a gravitational wave, ripples in the fabric of space-time, was detected and estimated to be the result of two black holes colliding 1.3BN light years from Earth.
    - Since then, four more gravitational waves were detected, one as this chapter went to press.
- **Collaboration**: The project depended on continuous collaboration between experimentalists who developed the equipment and theorists who defined what a signal from two black holes colliding would look like, let alone collaboration scientists, institutes, and countries.

6.3 Baylor-Watson: Cancer Drug Discovery

The Baylor-Watson drug discovery project (Spangler et. al., 2014) is a wonderful example of data-driven discovery and automatic hypothesis generation that discovered two novel kinases as potential sources for cancer drug development. These results that were determined to have a very high likelihood of success were developed in three months using IBM's Watson compared with the typical multi-year efforts that typically discover one candidate in two years.

- **Objective and phenomenon**: Discover kinases that regulate protein p53 to reduce or stem cancerous cell growth that have not yet been evaluated as a potential cancer drug.
- **Project**: Two years starting in 2012 between IBM Watson and the Baylor College of Medicine.
- **Equipment**: Watson as a data science platform; PubMed as data repository containing a corpus of 23M medical research articles.
- **Data:** 23M abstracts reduced to 240,00 papers on kinases reduced to 70,000 papers on kinases that regulate protein p53.
- **Hypothesis**: Some of 500 kinases in the corpus regulate p53 and have not yet been used for drugs.
- **AI Models / methods:** network analysis (Nagarajan, 2015) including textual analysis, graphical models of proteins and kinases, and similarity analysis.
- **Pipeline**: Explore, Interpret, and Analyze
    - **Explore:** Scan abstracts to select kinase papers using text signatures.
    - **Interpret:** Extract kinase entities from papers and build connected graph of similarity amongst kinases.

[28] http://www.ligo.org/

    - **Analyze**: Diffuse annotations over kinases to rank order the best candidates for further experimentation.
- **Data discovery and acquisition:** Textual analysis of PubMed.
- **Challenge:** Designing, developing and tuning models and methods to scan abstracts for relevant papers; to construct a graphical model of the relevant relationships, to select kinases that regulate p53.
- **Execution:** 3 months.
- **Results**: Two potential cancer drugs in 3 months versus 1 every 2 years (acceleration).
- **Validation**: The methods discovered 9 kinases of interest analyzing the corpus up to 2003; 7 of 9 were empirically verified in the period 2003-2013. This raised the probability that the remaining two that had not yet been verified clinically, were highly likely candidates.
- **Causality**: Work is underway to develop drugs that use the kinases to regulate p53 to stem or reduce cancerous cell growth.
- **Collaboration**: The project involved collaboration between genetic researchers, oncologists, experts in AI and natural language understanding, and computer scientists.

7 How important is collaboration in data science?

Data science is an inherently multidisciplinary activity, just as most human endeavors require knowledge, expertise, methods, and tools from multiple disciplines. Analyzing real world phenomena requires multidisciplinary approaches, e.g., how can you analyze the politics of a significant event without considering the economic factors (Brodie, 2015c)? Data science requires expertise from multiple disciplines, from the subject domain, statistics, AI, analytics, mathematics, computing, and many more. However, multidisciplinary collaboration is especially critical for success at this early time in the emergence of data science. Success and advancement in research and industry are typically based on competitive achievements of individual people or teams rather than on collaboration. While collaboration and multidisciplinary thinking are praised, they are seldom taught or practiced. Successful data science requires a behaviour change from competition to collaboration.

For disciplines required by scientific activities, there are well-established principles, methods, and tools from each discipline as well as how they are applied across scientific workflows. Collaboration was built into these mature disciplines and workflows years ago. In contrast, the principles, methods, and tools for each relevant discipline are just emerging for data science, as are methods of collaboration across workflows.

Currently, data science requires a data scientist to know the sources, conditions, and nature of the data to ensure that the domain specific model has the appropriate data. Rather than becoming a data expert the data scientist collaborates with a data expert. Rather than becoming an AI expert, a data scientist may need to collaborate with an AI expert to ensure the appropriate analytical methods are used. There can be as many as ten[29] disciplines involved in such an activity. Two current challenges in this regard are: 1) the shortage of data science-savvy experts, and 2) moving from a world of individual work to one of collaboration. Both challenges are being addressed by universities and institutes worldwide; however, the knowledge, as discussed above, and the teachers are themselves new to this game.

The need for collaboration on basic research and engineering on the fundamental building blocks of data science and data science infrastructures can be seen in a recent report from University of California, Berkeley researchers (Stoica, 2017). The report is a collaborative effort from experts from many domains – statistics, AI, data management, systems, security, data centers, distributed computing, and more.

Data science activities have emerged in most research labs in most universities and national research labs. Until 2017, many Harvard University departments had one or more groups conducting data science research and offered a myriad of data science degrees and certificates. In March 2017, the Harvard Data

[29] 10 is a somewhat arbitrary number chosen because most pipelines involve 5 to 10 expert tasks. The actual number of required disciplines varies significantly from simple analyses (e.g., reordering products for an online retailer) to very sophisticated (e.g., LIGO required analysis of ~1,000 sources of motion).

Science initiative[30] was established to coordinate the many activities. This pattern has repeated at over 120 major universities worldwide, resulting in over 150 DSRIs[31] being established since 2015 – themselves just emerging. The creation of over 150 DSRIs in approximately two years, most heavily funded by governments and by partner industrial organizations, is an indication of the belief in the potential of data science not just as a new discovery paradigm, but as a basis for business and economic growth.

Collaboration is an emerging challenge in data science not only at the scientific level but also at the strategic and organizational levels. Analysts report that most early industry big data deployments failed due to a lack of domain-business-analytics-IT collaboration (Forrester, 2015). Most of the over 150 DSRIs involve a grouping of departments or groups with an interest in data science, each in their own domain, into a higher level DSRI. A large example is the Fraunhofer Big Data Alliance[32], which in the above terminology would be a DSRI of DSRIs, describes itself as: "The Fraunhofer Big Data Alliance consists of 30 institutes bundling their cross-sector competencies. Their expertise ranges from market-oriented big data solutions for individual problems to the professional education of data scientists and big data specialists."

In principle, a DSRI would strive for higher-level, scientific and strategic goals, such as contributing to data science (i.e., the science underlying data science) in contrast with the contributions made in a specific domain by each partner organization. But how does the DSRI operate? How should it be organized so as to encourage collaboration and achieving higher-level goals?

While data science is inherently multi-disciplinary, hence collaborative, in nature, scientists and practitioners lack training in collaboration and are motivated to focus on their objectives and domain. Why would a bioinformaticist (bioinformatician) attempt to establish a data science method that goes beyond her requirements, especially as it requires an understanding of domains such as deep learning? Collaboration is also a significant organizational challenge specifically for the over 150 DSRIs that were formed as a federation of organizational units each of which conduct data science activities in different domains. Like the bioinformaticist, each organization has its own objectives, budget, and investments in funding and intellectual property. In such an environment, how does a DSRI establish strategic directions and set research objectives? One proposal is through a DSRI Chief Scientific Officer (Brodie, 2018b).

8 What is world-class data science research?

While many data science groups share a passion for data science, they do not share common data science components – principles, data, models, and methods; pipelines; data science infrastructures; and data infrastructures. This is understandable given the state of data science, and the research needs of the individual groups; however, to what extent are these groups pursuing data science, *per se*? This raises our original questions: *What is data science?* and *What is world-class data science research?* These questions are central to planning and directing data science research such as in DSRIs.

There are two types of data science research, domain specific contributions and contributions to the discipline of data science itself. Domain specific, world class data science research concerns applications of data science in specific domains resulting in domain-specific discoveries that are recognized in its domain as being world class. There are many compelling examples, as in section 6. To be considered data science, the research should adhere to the definition of data science, be based on some version of the data science method, use a data science pipeline, and utilize the components of data science. The data science components or the data science method, including scale, accelerating discovering, finding solutions that might not have been discovered otherwise, should be critical to achieving the result in comparison with other methods.

Equally or even more important, world class data science research should establish data science as a science or as a discipline with robust principles, data, models, and methods; pipelines; a data science method supported by robust data science infrastructures, and data infrastructures applicable to multiple domains. Such a contribution must be proven with appropriate applications of the first type. A wonderful

[30] https://datascience.harvard.edu/

[31] The DSRI list that I maintain by searching the web grows continuously - an excellent exercise for the reader.

[32] https://www.bigdata.fraunhofer.de/en.html

example of generalizing a domain-specific data science method is extending the network analysis method applied to some specific medical corpora used successfully in drug discovery (Spangler et. al., 2014) to domain-independent scientific discovery applied to arbitrary scientific corpora (Nagarajan, 2015). The original method was implemented in three stages, Exploration, Interpretation, and Analysis, using a tool call Knowledge Integration Toolkit (**KnIT**). Exploration involved lexical analysis and text mining of abstracts of the entire corpora up to 2003 (240,000) of medical literature mentioning kinases, a type of protein that governs cell growth, looking for proteins that govern p53, a tumor suppressor. This resulted in 70,000 papers to analyze further. Interpretation analyzed some of the papers to produce a model of each kinase and built a connected graph that represents the similarity relationship among kinases. The analysis phase identified and eliminated kinases that are not p53, ultimately resulting in discovering nine kinases with the desired properties. A retrospective search of the literature verified that seven of the nine were proven empirically to be tumor suppressors (candidates for cancer drugs) in papers published 2003-2013. This significantly raised the probability that the 2 remaining kinases were as yet undiscovered candidates for cancer drugs. These were world-class data science results and a magnificent example of analysis involving complexity beyond human cognition. First and foremost, the two kinases were accepted by the medical community as candidate tumor suppressors, i.e., published in medical journals. Second, the discovery was due to data science methods. Data science accelerated discovery since typically one such cancer drug candidate is found every two to three years; once the KnIT model was built the candidate kinases were discovered in approximately three months. The verification method, the retrospective analysis of cancer drug discovery 2003-2013 was brilliant. As with most data science analysis, the results were probabilistic, i.e., the nine candidate kinases were determined to likely candidates by the network model of the kinases, however, verification, or further confirmation, was established by a method outside data science altogether, i.e., discovered previously published results. The original analytical method that provided automated hypothesis generation (i.e., these kinases are similar) based on text mining of medical corpora concerning proteins was generalized to automated hypothesis generation based on text mining of any scientific corpora. While the first result was domain-specific, hence an application of data science, the extension of the domain-specific method to all scientific domains was *a contribution to the science of data science*. This is a higher level of world-class data science research.

The charter of every DSRI should include both domain-specific data science research and research to establish data science as a discipline. Since most DSRIs were formed from groups successfully practicing domain-specific data science, they are all striving for world class domain-specific data science. Without world class research in data science *per se*, it would be hard to argue that the DSRI contributes more than the sum of its parts. One might argue that lacking research into data science *per se* means that the DSRI has more of an organizational or marketing purpose than a research focus. The primary objective of a significant portion of the 150 DSRIs referenced above appears to be organizational, e.g., to bring together the various organizations that conduct data science. In contrast, in 2012 the Irish Government established Insight Center for Data Analytics as a national DSRI to conduct data science research and apply it in domains relevant to Ireland's future. In doing so it set objectives much higher than bringing together data science activities from its seven universities. The government of Ireland, through its funding agency, Science Foundation Ireland (SFI), continuously evaluates Insight on world class data science. This includes advancing data science principles, data, models, and methods and proving their value by achieving results in health and human performance, enterprises and services, smart communities and internet of things, and sustainability. More challenging, however, SFI requires that Insight contributes more than the sum of the parts, the individual units working on their own. This contributes to the science of data science by developing principles, data models, methods, pipelines, and infrastructure that is applicable to multiple domains.

9 Conclusions

Data science is an emerging paradigm with the primary advantage of accelerating discovery of correlations between variables at a scale and speed beyond human cognition and previous discovery paradigms. Data science differs paradigmatically from its predecessor scientific discovery paradigms that were designed to discover causality – *Why a phenomenon occurred* - in real contexts. Data science is

designed to discover correlations – *What phenomena may have or may occur* - in data purported to represent some real or imagined phenomenon. Unlike previous scientific discovery paradigms that were designed for scientific discovery and are now applied in many non-scientific domains, data science is applicable to any domain for which adequate data is available. Hence, the potential of broad applicability and accelerating discovery in any domain to rapidly reduce the search space for solutions holds remarkable potential for all fields. While already applicable and applied successfully in many domains, there are many challenges that must be addresses over the next decade as data science matures.

My decade-long experience in data science suggests that there are no compelling answers to the questions posed in this chapter. This is due in part to its recent emergence, it's almost unlimited breadth of applicability, and to its inherently multidisciplinary, collaborative nature.

To warrant the designation *data science*, this emerging paradigm, as a science, requires fundamental principles and techniques applicable to all relevant domains. Since most "data science" work is domain specific, often model- and method-specific, "data science" does not yet warrant the designation of a science. This is not a mere appeal for formalism. There are many challenges facing data science such as validating results thereby minimizing the risks of failures. The potential benefits of data science, e.g., in accelerating the discovery of cancer cures and solutions to global warming, warrant establishing rigorous, efficient data science principles and methods that could change our world for the better.

10 References

Braschler, M., Stadelmann, T. & Stockinger, K. (Eds.) (2018). "Applied Data Science - Lessons Learned for the Data-Driven Business", Berlin, Heidelberg: Springer, expected 2018

Brodie, M.L. (2014a) "*The First Law of Data Science: Do Umbrellas Cause Rain?*," *KDnuggets*, Jun. 2014.

Brodie, M.L. (2014b) "*Piketty Revisited: Improving Economics through Data Science – How Data Curation Can Enable More Faithful Data Science (In Much Less Time)*," *KDnuggets*, Oct. 2014.

Brodie, M.L. (2015a). Understanding Data Science: An Emerging Discipline for Data-Intensive Discovery, in Shannon Cutt (ed.), Getting Data Right: Tackling the Challenges of Big Data Volume and Variety, O'Reilly Media, Sebastopol, CA, USA, June 2015

Brodie, M.L. (2015b) Doubt and Verify: Data Science Power Tools, KDnuggets, July 2015. Republished on ODBMS.org.

Brodie, M.L. (2015c) *On Political Economy and Data Science: When A Discipline Is Not Enough*, KDnuggets, November 2015. Republished ODBMS.org November 20, 2015.

Brodie, M.L. (2018a) *Why understanding truth is important in Data Science*? KDnuggets, January 1, 2018. Republished Experfy.com, February 16, 2018.

Brodie, M.L. (2018b). On Developing Data Science, to appear in (Braschler, et. al. 2018)

Cambridge Mobile Telematics, (2018). Distraction 2018: Data from over 65 million trips shows that distracted driving is increasing, April 2, 2018.

Castanedo, F. (2015). Data Preparation in the Big Data Era: Best Practices for Data Integration, O'Reilly, August 2015.

Dasu, T. & Johnson, T. (2003) "Exploratory Data Mining and Cleaning," Wiley-IEEE (2003).

Data Science (2018), Opportunities to Transform Chemical Sciences and Engineering, A Chemical Sciences Roundtable Workshop, National Academies of Science, February 27-28, 2018

Demirkan, H. & Dal, B. (2014) The Data Economy: Why do so many analytics projects fail? Analytics Magazine, July/August 2014.

Dietterich, T. G. (2000, June). Ensemble methods in machine learning. In International workshop on multiple classifier systems (pp. 1-15). Springer, Berlin, Heidelberg.

Dingus, T. A., et. al. (2016). Driver crash risk factors and prevalence evaluation using naturalistic driving data. Proceedings of the National Academy of Sciences, 113(10), 2636–2641. http://doi.org/10.1073/pnas.1513271113

Duggan, J. & Brodie, M. L. (2015). "Hephaestus: Data Reuse for Accelerating Scientific Discovery," CIDR 2015, Jan. 2015.

Economist (April 2017). How Germany's Otto uses artificial intelligence, *The Economist*, April 12, 2017.

Economist (May 2017). The World's most valuable resource, *The Economist*, May 4, 2017.

Economist (January 2018) Many happy returns: new data reveal long-term investment trends, *The Economist*, January 6, 2018.

Economist (February 2018). Economists cannot avoid making value judgments: Lessons from the "repugnant" market for organs, The Economist, Feb 24, 2018.

Economist (March 2018). In algorithms we trust: How AI is spreading throughout the supply chain, The Economist, Mar 28, 2018

Eriksson, J., Girod, L., Hull, B., Newton, R., Madden, S., & Balakrishnan, H. (2008) The pothole patrol: using a mobile sensor network for road surface monitoring. In Proceedings of the 6th international conference on Mobile systems, applications, and services (MobiSys '08). ACM, New York, NY, USA.

Forrester (2015). Predictions 2016: The Path from Data to Action for Marketers: How Marketers Will Elevate Systems of Insight. Forrester Research, November 9, 2015

Forrester (2017). The Forrester Wave: Predictive Analytics and Machine Learning Solutions, Q1 2017, March 7, 2017.

Gartner G00310700 (2016) Survey Analysis: Big Data Investments Begin Tapering in 2016, Gartner, September 19, 2016.

Gartner G00301536 (2017). 2017 Magic Quadrant for Data Science Platforms, 14 February 2017.

Gartner G00326671 (2017). Critical Capabilities for Data Science Platforms, Gartner, June 7, 2017.

Gartner G00315888 (2017) Market Guide for Data Preparation, Gartner, 14 December 2017

Hey,T., Tansley, S. & Tolle,K. (Eds.) (2009) The Fourth Paradigm: Data-Intensive Scientific Discovery Edited by Microsoft Research, 2009

Jenkins, J. M., et. al. (2010). Caldwell, D. A., Chandrasekaran, H., Twicken, J. D., Bryson, S. T., Quintana, E. V., et al. (2010). Overview of the Kepler Science Processing Pipeline. The Astrophysical Journal Letters, 713(2), L87.

Liu J. T. (2012). "Shadow Theory, data model design for data integration," CoRR, vol. 1209, 2012. arXiv:1209.2647

Lohr, S. (2014) For Big-Data Scientists, 'Janitor Work' Is Key Hurdle to Insights, New York Times, August 17, 2014

Mayo, M. (2017) Data Preparation Tips, Tricks, and Tools: An Interview with the Insiders, KDnuggets, May 31, 2017

Nagarajan, M. et al. (2015). Predicting Future Scientific Discoveries Based on a Networked Analysis of the Past Literature. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '15). ACM, New York, NY, USA, 2019-2028.

NSF (2016). Realizing the Potential of Data Science, Final Report from the National Science Foundation Computer and Information Science and Engineering Advisory Committee Data Science Working Group, December 2016

Pearl, J. (2009a) Causality: Models, Reasoning, and Inference, New York: Cambridge University Press, 2009.

Pearl, J. (2009b). Epilogue: The Art and Science of Cause and Effect, In (Pearl, 2009a) pp. 401-428

Pearl, J. (2009c). "Causal inference in statistics: An overview," Statistics Surveys, 3:96--146, 2009.

Piketty, T. (2014). Capital in the 21st Century. The Belknap Press.

Press, G. (2016). Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says, Forbes, May 23, 2016

Reimsbach-Kounatze, C. (2015), "The Proliferation of "Big Data" and Implications for Official Statistics and Statistical Agencies: A Preliminary Analysis", OECD Digital Economy Papers, No. 245, OECD Publishing, Paris.DOI: http://dx.doi.org/10.1787/5js7t9wqzvg8-en

Spangler, S. et. al. (2014). Automated hypothesis generation based on mining scientific literature. In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '14). ACM, New York, NY, USA, 1877-1886.

Silver, D et. al. (2017). Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., et al. (2017). Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm. ArXiv E-Prints, cs.AI.

Singh, G. et. al. (2007). Optimizing Workflow Data Footprint Special issue of the Scientific Programming Journal dedicated to Dynamic Computational Workflows: Discovery, Optimisation and Scheduling, 2007.

Stoica,I, et. al. (2017). A Berkeley View of Systems Challenges for AI, Technical Report No. UCB/EECS-2017-159, October 16, 2017

Thakur, A. (2016). Approaching (Almost) Any Machine Learning Problem, The Official Blog of Kaggle.com, July 21, 2016.

Veeramachaneni, K. (2016) Why You're Not Getting Value from Your Data Science, Harvard Business Review, December 7, 2016.

Waller, M. A. and Fawcett, S. E. (2013), Data Science, Predictive Analytics, and Big Data: A Revolution That Will Transform Supply Chain Design and Management. J Bus Logist, 34: 77-84. doi:10.1111/jbl.12010

Winship, C., & Morgan, S. L. (1999). The Estimation of Causal Effects from Observational Data. Annu. Rev. Sociol., 25(1), 659–706. http://doi.org/10.1146/annurev.soc.25.1.659